Interpreting Deep Neural Networks

Bin Yu
Statistics and EECS, UC Berkeley

Workshop on Theory of Deep Learning: where next?
IAS, Oct. 17, 2019
ML/Stats Frontier: interpretation

EU's General Data Protection Regulation (GDPR) (2016) gives a “right” to explanation, and demands ML/Stats algorithms to be human interpretable

Image credit: https://christophm.github.io/interpretable-ml-book/
Examples of interpretation need

• FDA wants interpretation of DL algorithms for radiology

• *Interpretable gene interactions driving enhancer status for knock-out experiments

• *Stimuli to characterize a neuron

• *Phrases making a sentence negative
Interpretation is necessary in scientific ML

What is scientific ML?

- It uses machine learning for scientific research to extract, from data, discoveries, theory, and knowledge

- It builds scientific principles in machine learning algorithms

- It iterates between the above two steps

- Results are subject to scientific standards
What is interpretable ML (iML)?
(Murdoch, Singh, Kumbier, Abbasi-Asl, and Y., PNAS, 2019)
“Interpretable Machine Learning: Definitions, Methods and Applications”

https://arxiv.org/abs/1901.04592

“We define interpretable machine learning as the extraction of relevant knowledge from a machine-learning model concerning relationships either contained in data or learned by the model. Here, we view knowledge as being relevant if it provides insight for a particular audience into a chosen problem. These insights are often used to guide communication, actions, and discovery.”
iML-PDR in one figure

(P) Predictive accuracy  (D) Descriptive accuracy

Problem, Data, & Audience → Model → Post hoc analysis

(R) relevancy

R is key in the trade-off of P and D
iML through the PDR desiderata

• P- Predictive accuracy
  average (global) and point-wise (local)

• D- Descriptive accuracy: the degree to which an interpretation method objectively captures the relationships learned by machine learning models (both post-hoc and model-based methods can increase D)

• R- Relevancy: interpretation method is “relevant” if it provides insight for a particular audience into a chosen domain problem

Relevancy often plays a key role in determining the tradeoff between predictive and descriptive accuracy
There are cases where increasing D doesn’t decrease P.

Relevancy is often key

D vs P for model-based interpretability

Predictive accuracy

Descriptive accuracy

DNNs

Random forests

Logistic regression

Decision trees
Model-based interpretability

• Sparsity (e.g. sparse logistic regression for lung cancer prediction)

• Simulatability (e.g. decision tree for lung cancer prediction)

• Modularity (e.g. generalized additive models, layers in DL)

• Domain-based feature engineering (e.g. credit score)

• Model-based feature engineering (e.g. clustering and dimensionality reduction like PCA)
Post-hoc interpretability

• Data set level (global) interpretation (feature and interaction importance, statistical significance score, visualization)

• Prediction-level (local) interpretation (feature importance and alternatives)
Rest of the talk

Two post-hoc interpretation methods

• Project I: DeepTune (global) for neuroscience

• Project II: ACD (Agglomerative Contextual Decomposition) (local) for general DNN interpretation
Project I
The DeepTune framework for modeling and characterizing neurons in visual cortex area V4

Abbasi-Asl, Chen, Bloniarz, Oliver, Willmore, Gallant, and Y. (submitted, 2018)
https://www.biorxiv.org/content/early/2018/11/09/465534
Culmination of 3+ years of work

Reza Abbasi-Asl
Yuansi Chen
Adam Bloniarz

In collaboration with

Mike Oliver
Ben Willmore
Jack Gallant
Our approach to sML

“Embedded” students/postdocs work on site, in the wet lab

Seed scientific problem(s) → Generalization

Generalization: workflow, algorithms, theory
Interface between Neuroscience and Deep Learning

• Human visual cortex
  V4 is a difficult and elusive area

• Deep convolutional neural networks

[Image of brain diagram and neural network diagram]

http://cs231n.github.io/assets/nm1/neural_net2.jpeg
V1 decoded by Hubel and Wiesel (1959)

V1: orientation and location selectivity, and excitatory and inhibitory regions.

Nobel Prize in 1981
V4 has been probed by synthetic polar and hyperbolic gratings and complex shape stimulus

Gallant et al. 1993, 1996

Cartesian

Polar

Hyperbolic

David et al (2006)
V4 has been probed by synthetic convex and concave boundary stimuli

Pasupathy and Connor 1999, 2002

The stimuli were created by systematically combining convex and concave boundary elements.
Our data collection: 71 V4 neurons
(from the Gallant Lab at UC Berkeley)

Well-isolated visual neurons

Neuronal behavior is probed using sequences of natural images
Related works

Mairal et al (2013-, in prep): earlier work from us that uses sparse coding and SIFT to construct a two-layer NN with state-of-the-art predictive performance


Here we replicate their predictive results and aim at interpretation and understanding.
Questions to answer

1. How do we characterize V4 neurons?

   If we can characterize a neuron, we then know how to generate data-driven hypotheses.

2. How much do Convolutional Neural Networks (CNNs) resemble brain function?
DeepTune in a nutshell

Transfer predictive learning based on DNN+reg to derive 18 state-of-art prediction models for our V4 neurons (prediction).

System neuroscience insights into neurons through stable interpretation via DeepTune images of predictive models to suggest what V4 neurons do (stability).

As a result, we provide some support for resemblance of CNNs to primate brain, and generate image stimuli for closed-loop experiments.
Transfer learning...

**Step 1**
Training CNN

Images from ImageNet dataset

CNN used in Training for classification task

Labels of Images

**Step 2**
Feature Extraction And Fitting

Limited Images used in experiment

Early layers of trained CNN

Linear regression fitting via Ridge or Lasso

V4 neural activity

Prediction performance across different layers of CNN (AlexNet): N2 works well for V4
DeepTune image generation: Neuron 1

DeepTune Image(s):
Maximizing a (regularized) fitted model
Stable curve patterns across structurally compressed models

DeeTune image from full network

DeeTune images from compressed networks

Abbasi-Als and Y. (2017)
Top curve images from training set based on a model for neuron 1
Top curve images from test data set without models for Neuron 1
Stable predicted neuron activity from three deep nets +Lasso for a particular neuron
Dealing with multiple predictive models

CNN (e.g. AlexNet) + regression gives state-of-art prediction for V4 neurons – 18 such models

**Interpretation** via stability of DeepTune images over 18 models and several compressed models provides testable (prescriptive) characterizations of V4 neurons

We combat “**model-hacking**” via “**stability principle**”
It builds a unified platform to seek stability over data and algorithm perturbations.

Stability (aka robustness, invariance) is a minimum requirement for interpretability, reproducibility, and scientific hypothesis generation.
Neuron 1 seems a curve neuron and DeepTune images provide intervention stimuli

18 DeepTune images from 18 predictive models

<table>
<thead>
<tr>
<th></th>
<th>Ridge</th>
<th></th>
<th>LASSO</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Layer 2</td>
<td>Layer 3</td>
<td>Layer 4</td>
<td>Layer 2</td>
</tr>
<tr>
<td>AlexNet</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
</tr>
<tr>
<td>VGG</td>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
</tr>
<tr>
<td>GoogleNet</td>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
</tr>
</tbody>
</table>
Consensus DeepTune

- **Single model DeepTune**: Use gradient ascent to find stimuli that maximize one of the CNN+Regression model output.

- **Consensus DeepTune**: The models have to agree with each other to create a DeepTune pattern. (Stability)

\[ |\nabla f(x)| = \text{element-wise min} |\nabla f_i(x)| \]

\[ i = 1 \ldots \#\text{models} \]

Consensus Smooth DeepTune
Consensus DeepTune from 10 initializations

Neuron 1
Hierarchical clustering of "good" neurons through DeepTune Images on CNN feature space
Neuron 1: Predicted responses of cropped DeepTune

Small curve segments matter and responses compound due to convolution.

DeepTune images are suggestive of small curve segments being preferred stimuli of this neuron.
Neuron 1: regularity of spacing between curves seems an artifact of convolution filter size

<table>
<thead>
<tr>
<th>Ridge</th>
<th>LASSO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layer 2</td>
<td>Layer 3</td>
</tr>
<tr>
<td>AlexNet</td>
<td></td>
</tr>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
</tr>
</tbody>
</table>

18 DeepTune images from 18 predictive models
DeepTune images or parts are “verifiable” in closed-loop experiments

• cropped DeepTune images as stimulus images

• randomly cropped and combined images

• cropped images with varied sizes

Already done in
Viewing DeepTune from iML-PDR angle

• **Predictive accuracy:** state-of-the-art prediction performance on test data

• **Descriptive accuracy:** sparsity at last layer, modular, first layer Gabor – domain approved, some simulatability for DL part

• **Relevancy:** to the computational neuroscientists now (through peer review and talk feedback), neuroscientists (later), DL community (indirectly). Closed-loop experiments are very important steps forward
Computability of DeepTune

• Trained CNNs by others: stochastic gradient descent (SGD)

• Lasso/Ridge: gradient descent

• DeepTune: gradient ascent (descent)
A general approach
Three principles of data science: PCS

• **Predictability** for reality check

• **Computability**

• **Stability**: for the entire data science life cycle including data cleaning and data and algorithm/model perturbations

• **Transparent PCS documentation** (narratives and codes): “right” perturbations need to be argued for a particular goal
Examples of data perturbation

• Cross-validation partition, Bootstrap, Subsampling
• Adding small amount of noise to data
• Bootstrapping residuals in linear regression and liner time series models
• Block-bootstrap
• *Data perturbations through synthetic data such as mechanistic simulation PDE models
• *Adversarial examples in deep learning
• *Data under different environments/conditions (invariance)
• *Synthetic environments using the current data (stratification) (invariance relative to the stratification variable)
• Differential Privacy (DP)
• ...
Examples of model/algorithm perturbation

• Robust statistics models
• Semi-parametric models
• Lasso and Ridge models
• Different modes of a non-convex empirical minimization
• Different versions of Deep Learning algorithms
• Different kernel machines
• Sensitivity analysis of Bayesian modeling
• …
Causality evidence spectrum

Stable, replicable

Effect depends on the group

Stability implicit in causal inference: e.g. SUTVA

PCS workflow is relevant to causality:

Predictability + stability (aka robustness)

interpretability and hypothesis generation
Project II:

Agglomerative Contextual Decomposition (ACD)

(1) How can we get feature-interaction importance for a DNN model prediction in general? (ICLR 2018)

(2) How can we visualize these feature-interactions in an understandable way? (ICLR, 2019)

(3) How can we use the importance scores and prior info to debias algorithms? (submitted, 2019)
Previous work (post-hoc interpretation)

- gradient-based methods
  - LIME
  - Integrated Gradients (IG)
    - Ribeiro et al. (2016)
    - Sundarajan et al. (2017)

- contribution-based
  - Occlusion / saliency maps
    - Dabkowi & Gal (2017)
  - SHAP
    - Lundberg & Lee (2017)
An example from sentiment analysis

• Binary sentiment analysis with standard LSTM

The movie was good

LSTM

Positive

Explanation
Word importance scores can’t capture compositionality

3 pieces of information

3
1. not

2. good

3. not good

? → not good_1

not good_2

? → 2 outputs
CD: Contextual Decomposition

• Given a LSTM with weights, CD gives a prediction-level score for each phrase to “explain” the prediction

\[ \text{LSTM}(w_1, ..., w_T) = \text{SoftMax}(\gamma_T + \alpha_T) \]

• \( \gamma_T \) corresponds to contributions solely from the phrase, \( \alpha_T \) other factors

The movie was **not good**

Neg/ive
### Example: Sentiment Classification

<table>
<thead>
<tr>
<th>Occlusion</th>
<th>used</th>
<th>to</th>
<th>be</th>
<th>my</th>
<th>favorite</th>
<th>not</th>
<th>worth</th>
<th>the</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Integrated Gradients</td>
<td>used</td>
<td>to</td>
<td>be</td>
<td>my</td>
<td>favorite</td>
<td>not</td>
<td>worth</td>
<td>the</td>
<td>time</td>
</tr>
<tr>
<td>CD</td>
<td>used</td>
<td>to</td>
<td>be</td>
<td>my</td>
<td>favorite</td>
<td>not</td>
<td>worth</td>
<td>the</td>
<td>time</td>
</tr>
</tbody>
</table>

**Legend**
- Very Negative
- Negative
- Neutral
- Positive
- Very Positive
CD qualitatively picks up the correct regions.

<table>
<thead>
<tr>
<th>Image</th>
<th>CD</th>
<th>Occlusion</th>
<th>IG</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRT screen</td>
<td>![CD Map]</td>
<td>![Occlusion Map]</td>
<td>![IG Map]</td>
</tr>
<tr>
<td>Green mamba</td>
<td>![CD Map]</td>
<td>![Occlusion Map]</td>
<td>![IG Map]</td>
</tr>
<tr>
<td>Trash can</td>
<td>![CD Map]</td>
<td>![Occlusion Map]</td>
<td>![IG Map]</td>
</tr>
<tr>
<td>Crane</td>
<td>![CD Map]</td>
<td>![Occlusion Map]</td>
<td>![IG Map]</td>
</tr>
</tbody>
</table>
Agglomerative Contextual Decomposition (ACD)

Proc. ICLR

ACD is a hierarchical clustering algorithm with visualization, where the joining metric is CD scores
iML-PDR View of ACD

**Predictive accuracy:** interprets a trained model and does not change its predictive accuracy

**Descriptive accuracy:** allows for descriptions in terms of any subset of the feature space

**Relevancy:** to machine learning developers (to identify bias, perform sanity checks, and deal with interactions) and to the end users (to build trust, make the prediction process more transparent)
DNN Prediction

negative

not very good

DNN

ACD Interpretation

not very good

very good

not

very

good

Positive

Negative
a great ensemble cast can’t lift this heartfelt enterprise out of the familiar.
prediction: puck

skates are important

colors indicate different clusters

puck is important
Human experiments

Telling a good model from a “bad” one using only interpretations

Whether Interpretation instills trust or not
Improving models by regularizing ACD explanations

Rieger, Singh, Murdoch, Y. (2019). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge

In submission

CD/ACD code: github.com/csinvac/acd
Penalizing attributions for NLP (Liu & Avci, 2019)
Gradient saliency makes more sense (brighter = more saliency)

Test F1:

- Unregularized: 0.57
- Regularized: 0.62
Summary

- Interpretation is desirable for scientific machine learning and bias identification.
- It needs stability as a pre-requisite and implicitly depends also on predictability and computability – hence it needs PCS.
- Our iML framework: PDR
- Two interpretation methods: DeepTune and ACD
- On-going:
  - more empirical studies in the context of domain problems

CD/ACD code: github.com/csinvac/adc
Thanks to my group members and grants

Goal: quality research which is often slow

ARO and ONR
Paper links

1.* Three principles of data science: predictability, computability and stability (PCS) (Y. and K. Kumbier, 2019)
https://arxiv.org/abs/1901.08152

2*. Interpretable machine learning: definitions, methods and applications
https://arxiv.org/abs/1901.04592
Thank You!

Coming (2021?) ...

Data Science in Action: A Book
Bin Yu and Rebecca Barter
1Department of Statistics, UC Berkeley
2Department of Electrical Engineering and Computer Science, UC Berkeley

What skills do we teach?

Data Science In Action (DSIA) will teach the critical thinking, analytic, and communication skills required to effectively formulate problems and find reliable and trustworthy solutions. The primary skills taught are:

- Critical thinking
  - Formulate answerable questions using the data available
  - Scrutinize analytic decisions made and subsequent results
  - Document analytic decisions
  - Appropriate common techniques to unfamiliar situations
- Technical skills
  - Data processing skills
    - Data cleaning
    - EDA (numerical and visual summaries)
  - Algorithmic skills
    - Dimensionality reduction
    - Clustering
    - Least Squares & ML
  - Stability-based inferences
    - Data
    - Inference
  - Trustworthiness statements
    - Perturbation Intervals
    - Causal Inference
- Communication
  - Visual communication
    - “Exploratory” versus “explanatory” visual and numeric data summaries
  - Written communication
    - Each chapter has an open-ended case study for which the reader is encouraged to prepare a written analytic report

The Data Science Lifecycle is an iterative process that takes the analyst from problem formulation, data cleaning, exploration, algorithmic analysis, and finally to obtaining a verifiable solution that can be used for future decision-making.

The PCS framework provides concrete techniques for finding evidence for the connections between the three realms. Predictability: if the patterns found in the original data also appear in withheld or new data, they are said to be predictable. If an analysis or algorithm finds predictable patterns, then these patterns are likely to be capturing real phenomena. Computability: algorithmic and data efficiency and scalability is essential to ensuring that the results and solutions (e.g. a predictive algorithm) can be applied to new data. Stability: minimum requirement for reproducibility. If results change in the presence of minor modifications of the data (e.g. via perturbations) or human analytic decisions, then there might not be a strong connection between the analysis/ algorithms and the reality that underlies the data.

Intended Audience

Anyone who wants to learn the intuition and critical thinking skills to become a data scientist or work with data scientists. Neither a mathematical nor a coding background is required.

DSIA could form the basis of a semester- or multi-semester-long introductory data science university course, either as an upper-division undergraduate or early graduate-level course.
Contribution based methods could be problematic

How important is this region?

Zero background

Zero foreground

score

score
How important is this region?

Decompose each layer

Repeat...

CD Score
How does ACD work for a given layer?

- **Linear layer**: apply the linear weight to each part, and split the bias proportionally
- **Maxpool layer**: apply the maxpool layer to the combined image (relevant + irrelevant), take the max indexes, then use them to index the relevant / irrelevant parts separately
- **ReLU**: for the relevant part, apply the relu to the relevant part, for the irrelevant part apply the relu to both then subtract the relu of the relevant part
- **Quite general**: works for nearly any layer
$\beta = \text{relevant, } \gamma = \text{irrelevant, } i = \text{layer index}$

Linear/conv:

$$\beta_i = W \beta_{i-1} + \frac{|W \beta_{i-1}|}{|W \beta_{i-1}| + |W \gamma_{i-1}|} \cdot b$$

$$\gamma_i = W \gamma_{i-1} + \frac{|W \gamma_{i-1}|}{|W \beta_{i-1}| + |W \gamma_{i-1}|} \cdot b$$

Maxpool:

$$\text{max idxs} = \arg\max_{idxs} [\text{maxpool}(\beta_{i-1} + \gamma_{i-1}; idxs)]$$

$$\beta_i = \beta_{i-1}[\text{max idxs}]$$

$$\gamma_i = \gamma_{i-1}[\text{max idxs}]$$

ReLU:

$$\beta_i = \text{ReLU}(\beta_{i-1})$$

$$\gamma_i = \text{ReLU}(\beta_{i-1} + \gamma_{i-1}) - \text{ReLU}(\beta_{i-1})$$
MNIST example
More examples
Papers and upcoming book

1.* Three principles of data science: predictability, computability and stability (PCS) (Y. and K. Kumbier, 2019)
https://arxiv.org/abs/1901.08152

2. Book on data science (Y. and R. Barter, 2019, in prep)

3*. Interpretable machine learning: definitions, methods and applications
https://arxiv.org/abs/1901.04592
Spare slides
Berkeley’s DS Intellectual and Organizational Vision

Summary of the 2016 Report by the Faculty Advisory Board of the Data Science Planning Initiative
Prepared: 19 August 2016
Cathryn Carson, FAB Chair

Contents
A. Rationale for action: Why Berkeley, why now
B. Recommendations
   1. Organizational form: Core and connections
   2. Faculty FTE: Campus-wide surge and strategic foci
   3. Fundraising pillar and revenue generation
C. Situational challenges and next steps
D. The Faculty Advisory Board

Data8 Spring19 – 1500 students

CS/Stat Faculty co-creating and co-teaching data8.org and ds100.org

DS Interim Dean: D. Culler

New DS Major, Fall 2018

Div. of Data Science and Information headed by an Associate Provost (open search)

Data100 Spring19: 1,100 students
Thank you!

Safe and Green DS/AI

Image credit: https://www.ai-expo.net/drones-in-artificial-intelligence-are-they-safe/
Next steps for sML with empirical rigor

PCS (workflow and documentation) with iML-PDR is a step forward towards sML with empirical rigor. Moving forward, we need

• Consensus on evaluating empirical rigor in sML

• Consensus on standards when data results from scientific machine learning become knowledge

• Consensus on the process:
  debate among authors, peer reviews, follow-up experiments, ...

Fewer, and high quality papers would be a big help to sML and also to young researcher’s intellectual development
Parting thoughts:
engage in interdisciplinary research through people

• Broad interests and curiosity prepare for opportunities to arrive

• How do I know which opportunities to take on?

  If I like the people and they are good scientists, nothing could go wrong – in the worst case, I pick up some interesting science and have fun interacting…

  Many interactions do not lead to papers…